OXFORD

Measuring intramolecular connectivity in long RNA molecules using two-dimensional DNA patch–probe arrays

Timothy K. Chiang [1], Ofer Kimchi [2], Herman K. Dhaliwal[3], Daniel A. Villarreal[3],
Fernando F. Vasquez[3], Vinothan N. Manoharan [1,4], Michael P. Brenner[1,4], Rees F. Garmann [3,5,*]

[1]Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, United States
[2]Lewis–Sigler Institute, Princeton University, Princeton, NJ 08544, United States
[3]Department of Chemistry and Biochemistry, San Diego State University, San Diego, CA 92182, United States
[4]Department of Physics, Harvard University, Cambrigde, MA 02138, United States
[5]Viral Information Institute, San Diego State University, San Diego, CA 92182, United States

*To whom correspondence should be addressed. Email: rgarmann@sdsu.edu

Abstract

We describe a DNA-array-based method to infer intramolecular connections in a population of RNA molecules *in vitro*. First we add DNA oligonucleotide "patches" that perturb the RNA connections, and then we use a microarray containing a complete set of DNA oligonucleotide "probes" to record where perturbations occur. The pattern of perturbations reveals couplings between regions of the RNA sequence, from which we infer connections as well as their prevalences in the population, without reference to folding models. We validate this patch–probe method using the 1058-nucleotide RNA genome of satellite tobacco mosaic virus (STMV), which has been shown to have multiple long-range connections. Our results not only indicate long-range connections that agree with previous structures but also reveal the prevalence of competing connections. Together, these results suggest that multiple structures with different connectivity coexist in solution. Furthermore, we show that the prevalence of certain connections changes when pseudouridine, an important component of natural and synthetic RNAs, is substituted for uridine in STMV RNA, and that the connectivity of STMV minus strands is qualitatively distinct from that of plus strands. Finally, we use a simplified version of the method to validate a predicted 317-nucleotide connection within the 3569-nucleotide RNA genome of bacteriophage MS2.

Graphical abstract
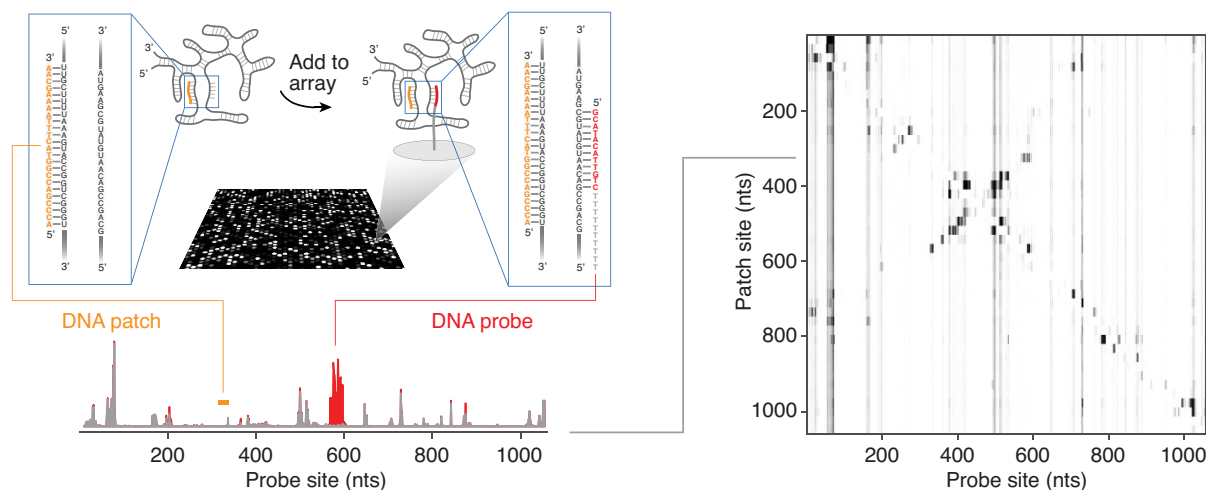
Introduction

Long sequences of ribonucleotides (RNA molecules containing a thousand nucleotides or more) not only can carry genetic information, but also can adopt functional structures that catalyze reactions and regulate cellular pathways. Their secondary structures span a range of scales: nucleotides nearby in the sequence can form local structures such as hairpins, and nucleotides separated by large distances along the sequence (hundreds to thousands of nucleotides) can pair to form long-range connections (Supplementary Fig. S1). These long-range

connections can significantly alter the overall size and shape of an RNA molecule [1–3] and the accessibility of its local structures [4]. Furthermore, a population of RNA molecules can have a distribution of competing long-range connections that might reflect different biologically relevant conformations [5–8].

While local structures in RNA molecules can be determined using a variety of methods, determining long-range connectivity and its variations is more challenging. Direct structural techniques, such as X-ray crystallography [9] and cryo-electron microscopy [10], are not well suited to long and heterogeneous RNA molecules. And indirect techniques, such as traditional chemical probing by SHAPE [11] and DMS [12], measure the conformational flexibility of each nucleotide in the sequence, which is related to the probability that each nucleotide is connected to another in the sequence but does not directly reveal the endpoint of the connection or if competing connections are present [13]. Thermodynamic folding models can be used to infer this missing information [14–16], but uncertainties in the model parameters affect the accuracy of detecting long-range connections [17].

Multidimensional probing techniques [18–25] address some of these challenges by more directly measuring connections between nucleotides. These techniques involve perturbing an RNA molecule at specific points and then detecting the effects of the perturbations elsewhere. For example, mutate-and-map [18], one of the earliest multidimensional probing methods, involves introducing point mutations into the sequence and then measuring corresponding changes in the conformational flexibility of the other nucleotides by chemical probing. Alternatively, PAIR-MaP [23], a more recent method, uses small molecule chemical probes to both perturb the RNA structure and detect changes in its flexibility. While these techniques can detect distributions of connections, the range of connections depends on the sequencing method used to measure the probing signals, with the upper bound set by the depth and length of the sequencing reads.

Proximity ligation techniques [26–31], such as PARIS [27], avoid read-length limitations by covalently linking connected nucleotides and then detecting the linked segments by sequencing. The protocols involve cross-link formation, fragmentation, enrichment of the linked fragments, ligation, removal of the cross-links, reverse transcription, PCR amplification, and high-throughput sequencing, followed by computational analysis of the sequencing data. While some research groups have successfully adopted these protocols, other groups may not have access to—or expertise in—the required techniques. Therefore, we aimed to develop a simpler method.

Our method of determining RNA connectivity is based on DNA probing, a technique in which RNA secondary structure is inferred from how strongly RNA molecules bind to complementary DNA oligonucleotides [32–38]. In contrast to the traditional DNA probing technique, developed over 50 years ago by Uhlenbeck *et al.* [32], we take a multidimensional approach similar to that described by Kaplinski *et al.* [39]. First we bind DNA oligonucleotide "patches" to specific regions of the RNA molecule to perturb the intramolecular connections [40]. Then we determine whether and where perturbations occur by measuring the binding of the patched RNA molecules to DNA oligonucleotide "probes" contained on a microarray [41]. The exceptional specificity of oligonucleotide hybridization enables us to perturb the RNA at many points and read out the perturbations in parallel on a single array, without covalent modification, reverse transcription, amplification, sequencing, or folding models. From the pattern of perturbations, we can infer long-range connections and their prevalences in a population of long RNA molecules.

Materials and methods

Buffers

Hybridization buffer: 50 mM Tris–HCl, pH 7.0; 1 M NaCl; 1 mM ethylenediaminetetraacetic acid (EDTA); 0.5% Tween 20. TE buffer: 10 mM Tris–HCl, pH 7.0; 1 mM EDTA. TAE buffer: 40 mM Tris–HCl, pH 8.3; 20 mM acetic acid; 1 mM EDTA. Note these buffers do not contain Mg, which is known to affect RNA folding. Although our method is compatible with Mg, we omit it to be consistent with previous measurements by Athavale *et al.* [42]

Preparation of fluorescently labeled RNA

We prepare fluorescently labeled RNA by *in vitro* transcription, using DNA templates derived from plasmids containing the RNA sequences (pSTMV [42], a gift from Steve Harvey, University of Pennsylvania; pSTMVm synthesized by Twist Bioscience; and pMS27 [43], a gift from David Peabody, University of New Mexico). The plasmids contain a T7 promoter sequence upstream of the RNA sequence, and a restriction site downstream of the RNA sequence. To generate a linear template, we digest the STMV (satellite tobacco mosaic virus)-containing plasmid with HindIII (New England Biolabs) and the MS2-containing plasmid with SmaI (New England Biolabs). We purify the template by acid phenol–chloroform extraction and ethanol precipitation, and resuspend the template in molecular biology grade water. The sequence of the template is verified by Sanger sequencing (Genewiz).

We transcribe STMV and MS2 RNA from the linear DNA templates using a TranscriptAid T7 High Yield Transcription Kit (Thermo Fisher). A small amount of Alexa Fluor 546 UTP (Thermo Fisher) is added to the transcription reaction (Supplementary Table S1), such that the final RNA transcripts contain roughly one dye per transcript, as measured by UV–Vis spectrophotometry (Supplementary Fig. S2). To prepare STMV RNA containing pseudouridine (Ψ), we replace the nonfluorescent UTP in the transcription reaction with ΨTP (TriLink BioTechnologies).

Following transcription, we digest the DNA template with DNase I (New England Biolabs) and purify the RNA transcripts using an RNeasy mini kit (Qiagen), eluting in TE buffer. Then, we wash the RNA five times with additional TE buffer using a 0.5-ml 100-kDa centrifugal filter (MilliporeSigma). The purified transcripts migrate as a single band both in a native 1% agarose gel prepared in TAE buffer and in a denaturing 1% agarose gel prepared in TAE buffer with 7 M urea, signifying that the transcripts are full-length and not degraded (Supplementary Fig. S2). Finally, we dilute the RNA to a concentration of 2 μM and store it at −80°C prior to use.

Sequence of the RNA transcripts

The sequences of the STMV RNA and MS2 RNA transcripts are listed in Supplementary data.

Automating the experimental design process

To facilitate the planning and execution of a patch–probe experiment, we have included a Jupyter Notebook that automates the experimental design process (Supplementary data). The user inputs the RNA sequence, patch length, probe length, and probe spacing, and the notebook generates a set of patch oligos and microarray design files. The design files include sequences and (x, y)-positions for all probe oligos and controls, in a format that is ready for direct submission to Agilent for array fabrication. The output can be adapted for use with other microarray manufacturers. See Supplementary data and Supplementary Fig. S3 for a description of how the patch length and probe spacing affect the number of microarrays required to map the connectivity of RNA molecules of varying length.

Design of the DNA patches

We design the DNA patches to be complementary to the STMV genome and to collectively tile its primary sequence. We use 44 patch oligonucleotides of length 24 nt that tile the STMV genome from nucleotides 1 to 1056. The patch sequences are listed in Supplementary Table S2. We also design a 20-nt DNA patch that is complementary to MS2 RNA at nucleotides 1427–1446 with the following sequence (5′–3′): AGTTAGAGCTGATCCATTCA.

Design of the DNA microarray and probes

The DNA microarrays used in our experiments are manufactured by Agilent Technologies (Supplier Item G4860A, SurePrint G3 Custom GE 1x1M Microarray). The arrays are prepared on 76 mm × 25 mm glass slides. Each slide contains 974 016 features arranged in a hexagonal grid with 1068 rows and 912 columns. Each feature is a circular spot with a nominal diameter of 30 mm in which many identical DNA oligonucleotides are tethered by their 3′-ends to the glass surface. Each oligonucleotide is 60 nt long.

We design two sets of DNA probes for the microarray: 12-nt probes and 24-nt probes. The 5′-ends of the probes contain either 12 or 24 nt that are complementary to the STMV genome, and the 3′-ends are poly-T spacers. We design each set of probes such that the complementary regions cover the entire STMV genome: there are 1047 12-nt probes and 1035 24-nt probes. The probes are numbered according to the first nucleotide in the STMV sequence that they are complementary to. As examples, the first and last three 12- and 24-nt probes are shown in Supplementary Table S3. The complete list of probes is given in Supplementary data.

To perform multiple patch–probe experiments on the microarray, we divide the array into identical subarrays that contain three replicates of each probe. Each subarray has 149 rows and 43 columns, for a total of 6407 spots. Of these spots, $(3 \times 1047 + 3 \times 1035 =)$ 6246 are assigned to the 12- and 24-nt probes and their replicates. The remaining spots are assigned to various control oligonucleotides, including poly-T oligonucleotides, poly-A oligonucleotides, and oligonucleotides that are complementary to the start and end of the transcript but only partially complementary to the STMV genome. We position the poly-T and poly-A controls in the corners of the subarray, and assign random positions to the probes and their replicates, as well as the remaining controls. Because the nominal spacing between rows is 0.018 mm and between columns is 0.063 mm, the subarrays are roughly square, with a height and width of 2.73 mm (Supplementary Fig. S4). We arrange the subarrays in a square grid across the microarray. The grid has 21 complete columns and 7 complete rows, for a total of 147 complete subarrays (Supplementary Fig. S4). The position of every spot and its corresponding probe sequence are provided as an XML file in Supplementary data.

Design of the microarray gasket

We design a gasket to distribute the RNA samples to different subarrays. The gasket holds at least 10 μl of sample volume and covers an area on the microarray that is greater than that of the subarray. Specifically, the wells cover 3.2 mm × 3.2 mm, with a separation of 1 mm between wells, and a separation of 0.5 mm from the edge of the array (Supplementary Fig. S5). A grid of wells consisting of 4 rows and 12 columns could fit on a single array.

A negative mold of the gasket is designed using FreeCAD and printed using a Formlabs Form 3 3D printer. We cast the gaskets in the negative mold using polydimethylsiloxane (PDMS). A mixture of 10 parts DOWSIL 184 silicone elastomer base (Dow) and 1 part DOWSIL 184 silicone elastomer curing agent (Dow) is mixed vigorously, degassed under vacuum, and then poured into the mold. We cure the PDMS by incubating at 65°C overnight.

Design of the hybridization clamp

We design a hybridization clamp to hold together the microarray slide and PDMS gasket (Supplementary Fig. S6). The clamp is designed using FreeCAD and printed using a Formlabs Form 2 3D printer. We fasten the clamp with four 1″-binder clips (ACCO) (Supplementary Fig. S7). The handles of the binder clips are removed after assembly.

Hybridization of the RNA and the DNA patches

In parallel, we mix the STMV RNA transcripts and each DNA patch in a 1:1 molar ratio in hybridization buffer such that the final concentration of RNA (and patch) is 10 nM and the final volume is 20 μl. We also add 24-nt poly-T DNA oligomers to a final concentration of roughly 1 μM, which reduces nonspecific binding of the RNA to the surfaces of our sample chamber, but does not significantly change the microarray binding intensities (Supplementary Fig. S8). Before adding the RNA–DNA mixtures to the microarray, the mixtures are heated to 90°C and then cooled to 4°C at a rate of −1°C/s. This heating and cooling procedure breaks up RNA aggregates that can form during *in vitro* transcription and also helps drive hybridization of the RNA and the DNA patches. As a control experiment, we perform the hybridization without heating and cooling, as described in the "RNA-patch hybridization without heating" section.

To hybridize DNA to MS2 RNA, we mix MS2 RNA and the DNA patch in a 1:1 molar ratio in hybridization buffer such that the final concentration of RNA and patch is 100 nM. Tween 20 (Thermo Fisher) is added to a final concentration of 0.2% to reduce nonspecific binding of the RNA to the surfaces of the sample chamber.

Hybridization of the RNA and the DNA microarray probes

We add the STMV RNA–DNA mixtures to the microarray using a custom-built gasket. The gasket is made of PDMS and contains a 4 × 12 square grid of 48 wells. Each well covers a 3.2 mm × 3.2 mm square, and the separation between wells is 1 mm (Supplementary Fig. S5). We load 10 μl of each RNA–DNA mixture into its own well and place the microarray slide DNA-side down on top of the gasket. At this point in the experiment, the mixtures are not in contact with the array. We then clamp the microarray and gasket together using a custom-built chamber (Supplementary Figs S6 and S7), invert the chamber, and spin it in a swinging bucket centrifuge at 1000 rcf for several seconds to bring the RNA–DNA mixtures in contact with the array. The microarray is incubated at 37°C for 100 min, washed for 1 min with hybridization buffer, and dried completely.

For MS2 RNA, we perform the hybridization and wash steps according to Agilent's Microarray-Based Gene Expression Analysis protocol, with minor modifications. We add 40 μl of the RNA–DNA mixtures to a SurePrint G3 Custom GE 8x60K Microarray (G4102A) using an Agilent gasket slide (G2534-60014). We assemble the gasket-microarray sandwich in a Hybridization Chamber (G2534A), and incubate the sample in a rotating oven at 37°C for 120 min. To wash the microarray, we prepare one container of Gene Expression Wash Buffer 1 (5188-5325) at room temperature and one container of Gene Expression Wash Buffer 2 (5188-5326) at 37°C. On a flat bench top, we disassemble the Hybridization Chamber, submerge the gasket and microarray sandwich in the first bath of Gene Expression Wash Buffer 1, and pry apart the gasket and microarray slide with forceps. The microarray is immediately transferred to the second bath containing Wash Buffer 2, washed with magnetic stirring for 5 min, and dried completely.

Imaging the microarray

We image the fluorescence of the microarray using an Agilent SureScan Microarray Scanner with the resolution set to 2 μm and the bit depth set to 20 (Supplementary data). The fluorescence integrated over each spot is determined using Agilent FeatureExtraction Software, using a nonstandard extraction protocol provided by the Agilent development team upon request (Supplementary data). We also determined integrated fluorescence for each spot using our own image analysis protocol written in MATLAB (Supplementary data). The prevalences inferred with these different feature extraction protocols are in agreement, as shown in Supplementary Fig. S27.

Compiling 1D and 2D binding spectra

We compile 1D and 2D binding spectra by averaging the integrated fluorescence over all spots corresponding to a given probe. For the 1D binding spectra (Fig. 1E and Supplementary Fig. S9), we average over spots from four sections of the microarray containing replicate samples of unpatched RNA. For the 2D binding spectra (Fig. 1F and Supplementary Fig. S9), each patch is added to its own section and we average over spots from that section only.

## Inferring the coupling signals and prevalences

A detailed account of the inference procedure is described in Supplementary data. Graphical depictions of the Bayesian models for inferring the signals and prevalences are shown in Supplementary Figs S10 and S11. Detailed results of the Markov chain Monte Carlo (MCMC) approach to infer the signals are shown in Supplementary Figs S12–S15 and in Supplementary Tables S4–S9. Detailed results of the MCMC approach to infer the prevalences are shown in Supplementary Figs S16 and S17 and in Supplementary Tables S10–S12. A comparison of inference results from two replicate experiments on the same RNA sequences is shown in Supplementary Fig. S18.

RNA–patch hybridization without heating

As a control experiment, we perform RNA–patch hybridization without heating and cooling by mixing 10-fold higher concentrations of RNA and patch (100 nM each) in hybridization buffer and incubating at 37°C for 1 h. Prior to patch hybridization, the RNA sample is heated and cooled by itself to break up aggregates and reproduce the thermal refolding process. Following patch hybridization but prior to adding the samples to the microarray, we dilute the concentration of RNA (and patch) back to 10 nM. This no-heat hybridization control gives similar results to the thermal hybridization experiments (Supplementary Figs S19–S22), suggesting that the no-heat protocol can be used to measure the connectivity of RNA molecules without exposing them to high temperatures.

Measuring patch affinities by gel electrophoresis

We validate the patch affinities inferred from the microarray data by comparing them to affinities measured in bulk using gel electrophoresis. In contrast to the microarray measurements, the bulk measurements are performed using DNA patches that are fluorescently labeled and STMV RNA molecules that are unlabeled. The 5′-ends of the bulk patches are the same as the microarray patches, but the 3′-ends contain an additional 6-T spacer sequence terminated by a fluorescein dye (3′ 6-FAM, Integrated DNA Technologies). The 6-T spacer is designed to reduce nucleobase quenching of the fluorescence signal that can occur when the fluorescently labeled patch binds to the RNA. To perform the bulk binding measurements, we mix together 300 nM of STMV RNA with 300 nM of fluorescently labeled patch in hybridization buffer. In one experiment, we heat the sample to 90°C, cool the sample to 4°C at a rate of -1°C/s, and then incubate the sample for 100 min at room temperature. This protocol corresponds to our normal patch–probe microarray experiment. In another experiment, we heat and cool the RNA molecules before adding the patches, and then incubate the mixture of RNA and patch for 100 min at room temperature. This protocol corresponds to our no-heat control. Following incubation, we perform native 1% agarose gel electrophoresis in TAE buffer and image the fluorescence of the gel (Supplementary Fig. S23). The slower-running fluorescence signal corresponds to bound patch, and the faster-running band corresponds to unbound patch. The patch affinity is calculated by dividing the bound signal by the total (bound plus unbound) signal (Supplementary Fig. S24). Comparisons of the microarray measurements and the bulk measurements, corrected for concentration according to the
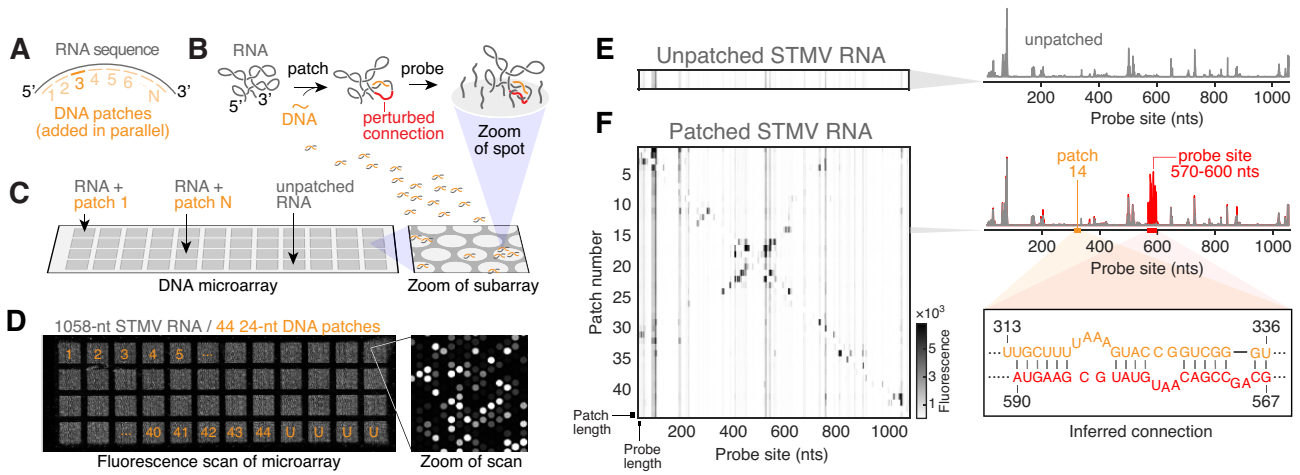
Figure 1. Basis of the patch–probe method. (**A**) Diagram of DNA patches that are complementary to the RNA. (**B**) Diagram of the patch–probe method, showing how a patch can perturb an intramolecular connection, affecting the binding of surface-tethered probes. (**C**) We add each patched RNA to its own section of a DNA microarray and add unpatched RNA to a separate subarray. The microarray consists of spots, each of which contains multiple copies of a probe tethered at its 3' end by a poly-T spacer. Within each section, every probe of a specified length that is complementary to the RNA appears in 3–12 spots. (**D**) A fluorescence micrograph of a microarray containing 12- and 24-nt probes. We determine binding by labeling the RNA and measuring the fluorescence of each spot, which is proportional to the number of bound RNA molecules. Here we show the integrated fluorescence in 48 sections, 44 of which contain patched STMV RNA and 4 of which contain unpatched ("U") STMV RNA. Each patch is 24 nt, and the 44 patches completely tile the STMV RNA sequence from nucleotides 1 to 1056. (**E**) Heatmap (left) and a bar plot (right) of the binding spectrum of unpatched RNA to 12-nt probes, as detailed in the "Materials and methods" section. This 1D measurement reveals a broad range of fluorescence magnitudes for the probes. (**F**) Heatmap of the binding spectrum of patched RNA to 12-nt probes (left). This 2D measurement consists of all the 1D patched RNA spectra sorted by patch number. The vertical lines correspond to those in the unpatched RNA heatmap. Features that stand out above this background reflect patch-induced perturbations in RNA connectivity. The binding spectrum for RNA hybridized to patch 14 (right, top) is shown in red, with the spectrum of unpatched RNA overlaid in gray. The binding site of patch 14 (nucleotides 313–336) is shown in yellow. The difference between the patched and unpatched spectra shows a large increase in binding at probes 200 nt downstream of the patch site, indicating a potential connection between these regions of the RNA sequence (right, bottom). See Supplementary Fig. S28 for a detailed diagram of these patch and probe interactions, and Supplementary Fig. S13 for binding spectra of 24-nt probes.

procedure described in Supplementary data, are shown in Supplementary Figs S25 and S26.

Identifying dominant couplings

For each patch, we identify the top 5 probes with the highest signal, S_{ij}. We define these patch–probe combinations as the dominant couplings. We plot each dominant coupling as a rectangle that spans the patch and probe binding sites.

Results

We apply the patch–probe method outlined in Fig. 1A–C to the 1058-nt RNA genome of STMV, a long-standing model system that has been studied using chemical probing [42, 44–46], computational modeling [47, 48], and direct imaging [44, 49]. We use fluorescently labeled STMV RNA (see the "Materials and methods" section), a series of 44 24-nt patches (see the "Materials and methods" section), and a single microarray containing all 12- and 24-nt probes that are complementary to the RNA (Fig. 1D and "Materials and methods" section). We discuss experiments with 12-nt probes throughout the main text of this paper, and we include experiments with 24-nt probes as Supplementary data.

Increases in probe binding after patching reveal intramolecular connections

The binding spectrum for unpatched RNA contains many peaks that, in principle, contain information about its connectivity (Fig. 1E), but such information is in practice difficult to extract from these peaks alone. The fundamental problem

is that 1D probing measurements do not directly determine which regions of the molecule are connected. Furthermore, any information about connectivity is convolved with variations in probe binding affinity (Supplementary Fig. S29).

We resolve this problem by examining how the patches affect the binding, yielding a 2D measurement. The 2D spectrum has a background of vertical lines that correspond to peaks in the unpatched spectrum and therefore reflect the affinity of probes to the unpatched RNA (Fig. 1E). However, several patch–probe combinations show features that stand out above the background of vertical lines (Fig. 1F). These features correspond to an increase in probe binding, which we expect to occur if a patch disrupts a connection to a probe site in some fraction of the RNA molecules. For example, upon addition of patch 14, we observe an increase in binding for probes 560–590 (Fig. 1F, right). We interpret the coupling between patch 14 and probes 560–590 as evidence of a connection in the RNA involving segments that contain the patch and probe binding sites. Such couplings contain more direct information about the connections than the background itself.

While the peaks in the raw 2D binding spectrum shown in Fig. 1F suggest a pattern of connections in the molecule, this pattern is obscured by the background. We therefore separate changes in binding from the background. Specifically, for each patch–probe combination, we define the coupling signal S_{ij} to be the change in how probe number i binds to RNA attached to patch number j, relative to how probe i attaches to unpatched RNA: $S_{ij} = I_{ij}/B_i - 1$, where I_{ij} is an estimate of the true patch–probe fluorescence and B_i is an estimate of the true background. We use a Bayesian approach to infer the signal S_{ij} because the approach allows us to quantify the un-
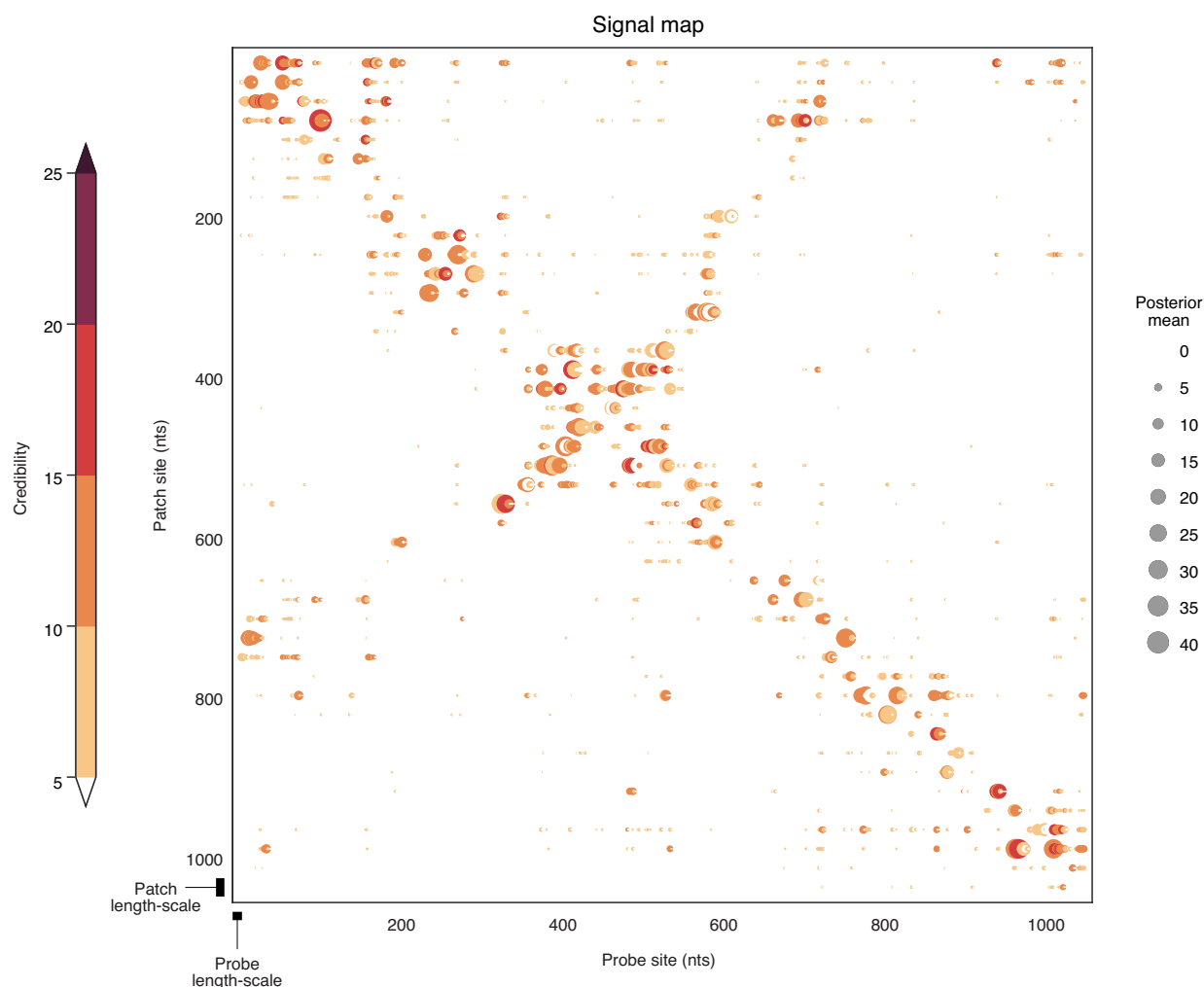
Figure 2. The patch–probe coupling signals suggest many possible connections in STMV RNA. The map of signals inferred from Fig. 1E, as described in Supplementary data. The size of each dot is proportional to the signal, and the color reflects the credibility, which we define as the ratio of the posterior mean to standard deviation. See Supplementary Fig. S13 for the map corresponding to 24-nt probes.

certainty on the signal using all of the measurements and to account for experimental effects such as outliers and variations in RNA amounts across subarrays. The statistical model and MCMC sampling scheme are detailed in Supplementary data.

The resulting signals and uncertainties, shown in Fig. 2, are independent of variations in probe affinity and in RNA amounts across the microarray (see Supplementary Fig. S12), allowing us to directly compare coupling strengths between a given patch site and all probe sites. We do not expect the map of the signals to be perfectly symmetric, for several reasons. First, there are variations in patch affinity. Second, the patches tile the RNA sequence end-to-end and need not align evenly with each connection in the underlying structure (Supplementary Fig. S30). Third, while thermodynamic folding models [16] suggest patch-induced RNA refolding is unlikely (Supplementary Fig. S31), it is possible that patch binding triggers large-scale structural changes in the RNA molecules that extend beyond individual pairwise connections. Nonetheless, the map shows a roughly symmetric pattern, with credible signals (as determined by the mean of the marginalized posterior of the signal divided by the standard deviation) occurring both near to and far from the diagonal. Signals near the diagonal, scattered from upper left to lower

right, reflect short-range connections. Signals that extend far away from the diagonal, such as the long central ridge that cuts perpendicular to it, reflect long-range connections.

Dominant patch–probe signals reveal consensus duplexes of STMV RNA

The signal map (Fig. 2) yields a large amount of information about the patch–probe couplings and the underlying connections. We focus first on the dominant couplings, corresponding to the probes with the largest signals for each patch (see the "Materials and methods" section). The map of dominant couplings, shown in Fig. 3A, shows many of the features that stand out in the raw binding spectrum of Fig. 1E.

Couplings near the center of Fig. 3A provide a valuable point of comparison between our measurements and previous structural studies. Three different SHAPE chemical probing studies [42, 44, 45], as well as direct measurements using atomic force microscopy [44] and cryo-electron microscopy [49], suggest that the central region of STMV RNA adopts a T-shaped domain containing three long-range connections: a 90-nt-long hairpin and a 270-nt-long hairpin branching from a 50-nt-long central duplex that connects regions over 470 nt apart (highlighted by ellipses in Fig. 3B).
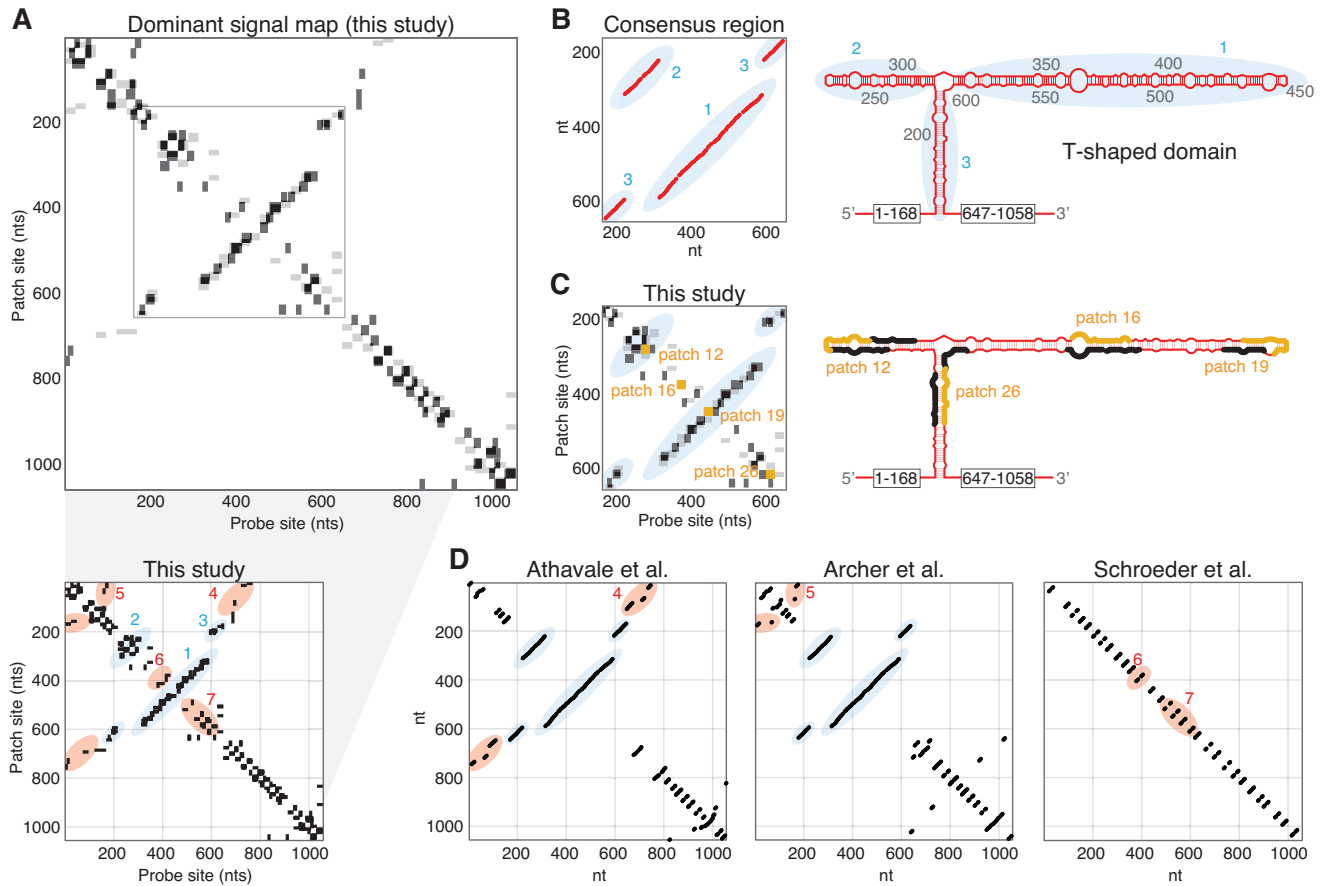
Figure 3. The dominant patch–probe signals are consistent with the consensus structure of STMV RNA and reveal competing structures in non-consensus regions. (**A**) The map of dominant couplings (top) shows the five probes with the largest signals for each patch. Each dominant coupling is plotted as a black rectangle spanning the patch and probe binding sites. We plot the transpose of the map in gray to highlight couplings that may be obscured by larger features induced by other patches. The gray box shows the central region of the RNA, which adopts the consensus T-shaped structure. At bottom is the same plot in which ellipses highlight three long-range connections in the central region (blue ellipses 1–3), two long-range connections outside of the central region (red ellipses 4 and 5), and two stretches of short-range connections (red ellipses 6 and 7). (**B**) A dot plot and a diagram of the consensus structure, with base pairs shown in red and three long-range connections highlighted by blue ellipses 1–3. (**C**) Plot of the central region [gray box in panel (A)] of the dominant coupling map, showing consistency with three long-range connections, highlighted by blue ellipses. The binding sites of four patches are shown in yellow. A diagram of the T-shaped domain of the consensus structure is shown in red, with the binding sites of the four patches highlighted in yellow, and the largest changes in probe binding shown in black. (**D**) Dot plots of the structures reported in previous chemical probing studies contain different sets of connections outside the consensus region. Some of these connections are seen in the set of dominant connections inferred from the patch–probe data: SHAPE probing of RNA transcribed *in vitro* by Athavale *et al.* [42] shows connections in ellipse 4; SHAPE probing of RNA extracted from virus particles by Archer *et al.* [44] shows connections in ellipse 5; DMS, kethoxal, and CMCT chemical probing and crystallographic analysis of RNA packaged in virus particles by Schroeder *et al.* [46] shows connections in ellipses 6 and 7. See Supplementary Fig. S32 for dominant couplings corresponding to 24-nt probes and arc plot representations of the data.

Although the resolution of our map, set by the patch and probe size, is coarser than that of a dot plot, our map reveals couplings that are consistent with all three long-range connections in the consensus T-shaped domain (Fig. 3C, left). Visualizing these couplings on top of a structural model of the T-domain further demonstrates the good agreement between our patch–probe measurements and the consensus structure (Fig. 3C, right).

Dominant patch–probe couplings detect competing connections

Having validated our results against the consensus structure, we now examine features outside the central T-shaped region. Here there is less consensus. Previous chemical probing studies by Athavale *et al.* [42], Archer *et al.* [44], and Schroeder *et al.* [46] show considerable differences in connectivity (Fig. 3D).

Some of these may be due to differences in the chemical probing protocol or the source of the RNA, and others might reflect differences in the folding models used to interpret the data. Both Athavale *et al.* and Archer *et al.* used thermodynamic folding models, but Archer *et al.* imposed a cutoff length of 600 nt for base-pair interactions. Schroeder *et al.* developed a cotranscriptional folding and assembly model with a short cutoff of 30 nt. The effects of these cutoffs are made clear by the dot plots in Fig. 3D. For example, the dot plot of the structure from Athavale *et al.* contains dots far from the diagonal that represent long-range connections, the longest of which connects nucleotides 12 and 746 (Fig. 3D, ellipse 4). These long-range connections are necessarily absent in the dot plots of the structures reported by Archer *et al.* and Schroeder *et al.* because they lie beyond the imposed cutoffs. Cutoffs can also affect the inference of shorter-range pairs by constraining the folding model. Such indirect effects could explain why mid-

range base pairs in the dot plot from Archer *et al.* (Fig. 3D, ellipse 5) are absent in the dot plot from Athavale *et al.*

With the above differences in mind, we compare our map of dominant couplings (Fig. 3A, bottom) to the previously reported structures (Fig. 3D). Outside of the central T-domain, we observe couplings far from the diagonal that are consistent with the longest-range connections in the structure reported by Athavale *et al.* (ellipse 4; see, for example, features between nucleotides 1–27 and 721–768 in Fig. 3A, bottom and between nucleotides 12–24 and 735–746 in Fig. 3D). We also observe slightly off-diagonal couplings consistent with mid-range connections in the structure reported by Archer *et al.* (ellipse 5), and several near-diagonal couplings consistent with local connections in the structure reported by Schroeder *et al.* (ellipses 6 and 7). Each of these connections is found in only one of the previous structures and not the others.

Some of these connected regions are unlikely to coexist within the same structure, suggesting that our measurements detect the presence of multiple structures within the population of RNA molecules. For example, we do not expect the long-range connected region shown in ellipse 4 to coexist with the shorter-range connected region shown in ellipse 5 in the same structure since these regions share a significant number of nucleotides. Consistent with our expectations, thermodynamic folding algorithms, such as RNAsubopt from the ViennaRNA software package [16], indicate that structures containing connected regions consistent with either ellipse 4 or ellipse 5 are energetically favorable, whereas structures containing connections consistent with both ellipses are not. Specifically, within a Boltzman-weighted population of 10^6 suboptimal structures predicted by RNAsubopt, 78% contain a stem consistent with either ellipse 4 or ellipse 5, but none contain stems consistent with both ellipses. Moreover, when we constrain the folding algorithm to include either stem, none of the 10^6 suboptimal structures contain the other. These results suggest that the connections measured in ellipses 4 and 5 reflect multiple structures within the population of RNA molecules. Furthermore, because these connections have markedly different ranges, their corresponding structures could have qualitatively different shapes and sizes.

That long RNA molecules might adopt a distribution of structures is not unexpected. Equilibrium base-pairing probabilities calculated using thermodynamic folding algorithms also predict a distribution of structures. Although we do not claim that our system is in equilibrium, we note that predictions of these algorithms are qualitatively consistent with our measurements. Specifically, RNAfold from the ViennaRNA software package [16] shows that nucleotides in the central region between 400 and 600 can form both short- and long-range connections (Supplementary Fig. S33), consistent with dominant couplings in ellipses 1–3 and ellipses 6 and 7. These results are also consistent with chemical probing studies by Larman *et al.* [45] that report evidence for competing structures over the same region of the sequence.

Normalizing by the patch affinities reveals additional RNA connections and their prevalence

Our analysis thus far has focused only on the dominant couplings. The signal map (Fig. 2) reveals many other less strong yet credible couplings that reflect additional connections. But because the coupling signals depend on both the RNA connectivity and the patch affinity, the signal map does not distinguish between, for example, couplings that arise be-

cause a large fraction of RNA molecules are connected at the patch site and those that arise because a smaller fraction are connected, but the patch is more efficient at disrupting those connections. To separate these effects and glean information about the prevalence of connections, we must account for variations in the patch affinities.

We infer the patch affinities directly from the microarray data, and specifically from measurements of probes binding to sites that overlap completely with patch sites (these data correspond to diagonal elements of the signal map). We use a linear model and a Bayesian inference scheme, as detailed in Supplementary data. Briefly, we assume that the coupling signal is linearly proportional to the patch affinity—the simplest assumption we can make—and infer a normalized signal $R_{ij} = S_{ij}/p_j$, where p_j is the patch affinity, or fraction of RNA molecules that are patched. We then rescale the normalized signal as $f_{ij} = R_{ij}/(1 + R_{ij})$ to produce values that lie between 0 and 1. Under a restrictive set of assumptions (see the "Equilibrium model for probe binding" section in Supplementary data), f_{ij} is the fraction of molecules in the population that contain a connection between probe site i and patch site j. But even when these assumptions are relaxed, we expect f_{ij} to scale monotonically (though not necessarily linearly) with how frequently the connection appears in the population, because the values are corrected for both patch and probe binding affinities. For this reason, we call f_{ij} the "prevalence" of connection ij. We can compare prevalences among all patch–probe combinations and even among different experiments.

The map of f_{ij} in Fig. 4A reveals many features with high prevalence in the population, including some features that have not been reported previously. For example, we observe a cluster of features with high prevalence far from the diagonal, which point to long-range connections spanning nearly the entire RNA sequence (Supplementary Fig. S34). These very long-range connections have not been reported in previous studies, but are predicted by RNAfold [16] to be present in the population of equilibrium structures (Supplementary Fig. S33). We also see isolated long-range features that are not part of a cluster and might reflect long-range pseudoknots (Supplementary Fig. S34). As with the dominant couplings, some features in the prevalence map point to competing connections that cannot appear in the same structure (Supplementary Fig. S34), providing further evidence that STMV RNA adopts multiple structures, at least under our experimental protocols.

The method reveals structural changes induced by modified nucleotides

Because the prevalences f_{ij} are corrected for patch and probe affinities, we can compare their values across experiments, enabling us to measure changes in RNA connectivity in response to changing conditions. To demonstrate this point, we apply the method to STMV RNA containing the modified nucleotide pseudouridine (Ψ) in place of normal uridine (U). Pseudouridine, an important component of natural RNA molecules [50] and synthetic RNA vaccines [51], forms stronger interactions with other nucleotides than uridine does [52]. These interactions are known to stabilize short duplexes [53], but their effect on the connectivity of long RNA molecules is not understood. By comparing patch–probe experiments on STMV RNA molecules with and without Ψ, we aim not only to demonstrate that the method can detect changes in connectivity, but also to measure which connections are affected by the modified nucleotides.
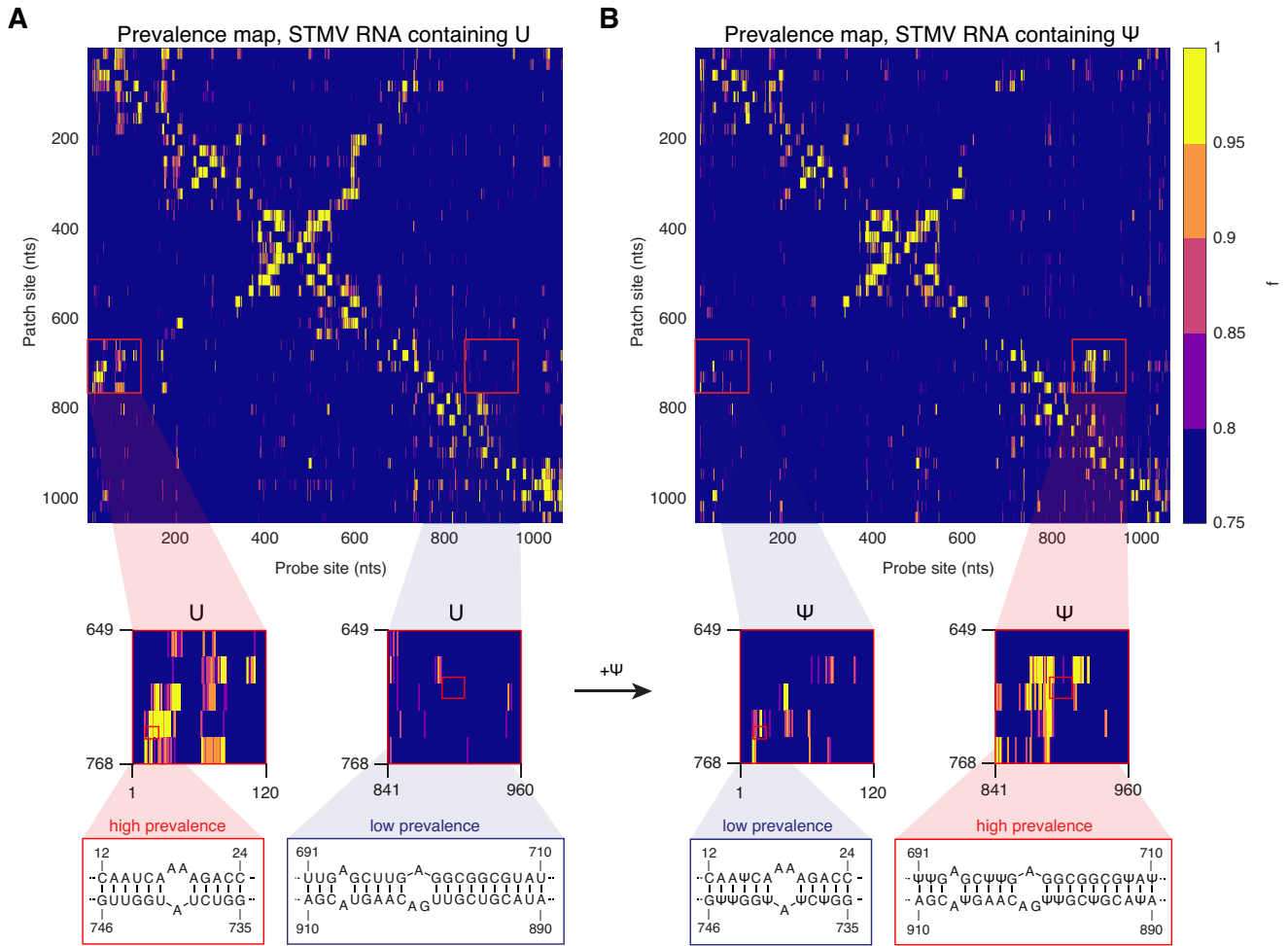
Figure 4. Normalized coupling signals provide information about the prevalence of connections in the population of STMV RNA molecules and enable comparisons between different experimental conditions. (**A**) The map of normalized coupling signals for STMV RNA, which is corrected for patch and probe affinities (see the "Materials and methods" section), provides a measure of the prevalence of connections in the population. (**B**) The map of normalized signals for STMV RNA containing pseudouridine (Ψ) reveals differences in the prevalence of certain connections relative to RNA containing normal uridine (U). Specifically, highly prevalent connections between patch sites at 649–768 nt and probe sites at 1–120 nt in U-containing RNA are less prevalent in Ψ-containing RNA. In contrast, low-prevalence connections between patch sites at 649–768 nt and probe sites at 1–120 nt in U-containing RNA show higher prevalence in Ψ-containing RNA. These differences are highlighted by red boxes. Partially complementary sequences in these regions suggest specific connections that may be present. See Supplementary Fig. S34 for additional plausible connections and Supplementary Figs S16 and S17 for details of the corrections for patch and probe affinities.

The prevalence map of Ψ-containing molecules (Fig. 4B) shows some features similar to those of unmodified STMV RNA, including the central ridge of the T-domain, but also shows some differences. In particular, we observe a decrease in one cluster of long-range couplings (Fig. 4A), and the appearance of a new cluster of shorter-range couplings (Fig. 4B). The long-range couplings (involving regions of the sequence between nucleotides 1–100 and 650–750) are consistent with connections reported by Athavale *et al.* (Fig. 3D, ellipse 4), and the shorter-range couplings (involving the region between nucleotides 670–920) are consistent with connections predicted by RNAfold [16] (Supplementary Fig. S33). Thus, the incorporation of Ψ appears to shift the connectivity of nucleotides 670–750 away from the longer-range connections toward the shorter-range connections, signifying a qualitative change in structure.

These changes occur in regions of the sequence that are thought to adopt functional structures. Downstream of nucleotide 700, the STMV sequence has high homology to the RNA of its helper virus, tobacco mosaic virus, which is known to fold into a functional transfer-RNA-like structure [54] flanked by multiple short-range pseudoknots [55]. By detecting changes in the connectivity of this region, our results suggest structural changes that might affect the functionality of STMV RNA. These results could be used to inform *in vivo* studies that directly test the effect of Ψ on the biological properties of STMV. Similar comparative studies on Ψ-containing molecules used in RNA vaccines could reveal structural changes that shed light on their unique biological properties, such as their enhanced translational capacity and reduced immunogenicity [51].

Analysis of the minus strand of the STMV genome

During replication, the STMV genome exists in two forms: a plus strand (discussed above) and a minus strand (its reverse complement). The genome enters and exits the host cell as a plus strand, but inside the cell, the RNA-dependent RNA polymerase (RdRp) of the helper virus synthesizes minus-strand copies that serve as templates for generating additional plus
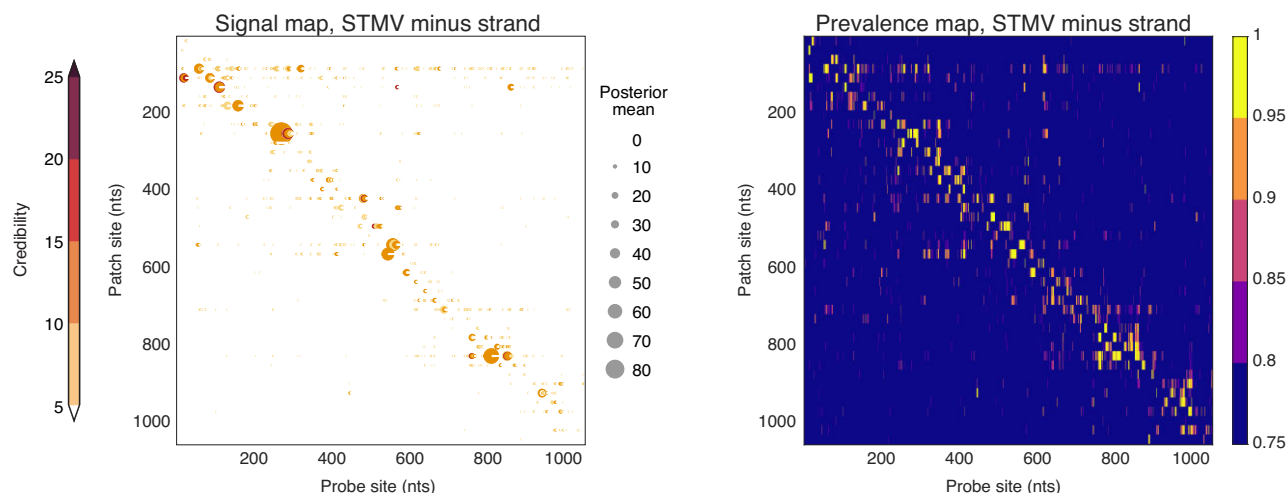
Figure 5. Signal and prevalence maps of the minus strand of the STMV genome reveal qualitative differences in connectivity compared to the plus strand. The signal map for the STMV RNA minus strand is shown on the left, and the prevalence map is shown on the right. These maps lack much of the off-diagonal signal observed with the plus strand, suggesting that the minus strand adopts qualitatively different structures with fewer long-range connections.

strands [56]. Although the minus strand is essential for viral replication and spread, its structure remains uncharacterized.

This gap in knowledge is not unique to STMV—structural studies of RNA viral genomes have focused primarily on plus strands, leaving minus strands largely unexplored. Because plus and minus strands can form identical G–C and A–U base pairs, one might expect their secondary structures to be similar. However, differences in G-U pairings, stacking interactions, and sequence-dependent folding kinetics could lead to important structural differences. The extent to which plus and minus strands adopt similar overall folds or diverge in functionally significant ways remains an open question.

Our method addresses this question by enabling direct comparison of plus- and minus-strand connectivity. The signal and prevalence maps for STMV RNA minus strands (Fig. 5) reveal striking differences from the maps of plus strands (Figs 2 and 4A). The minus-strand map contains fewer credible couplings than the plus-strand map (Fig. 5 versus Fig. 2), and the minus-strand couplings occur primarily along the diagonal, reflecting short-range connections. Notably, the off-diagonal couplings that define the T-shaped domain in plus strands are absent in minus strands (Fig. 5 versus Fig. 4A).

These structural differences could reflect distinct functional roles for the plus and minus strands. The central T-shaped domain of the plus strand is thought to provide stability against host cell nucleases [42] by adopting linear arrays of duplexes akin to those found in certain viroids [57]. Such stability could be critical during the early stages of STMV infection, in which recently uncoated plus strands must persist until the helper virus RdRp initiates replication. In contrast, minus strands are present only during active replication, reducing the selective pressure for nuclease resistance. The absence of viroid-like connectivity in minus strands is consistent with this interpretation.

Furthermore, X-ray crystallography studies of STMV show ordered duplexes of RNA bound to the interior capsid surface [58]. These RNA duplexes are thought to stabilize the capsid and facilitate genome packaging [59]. Since only one of the plus and minus strands must be encapsidated, the virus need not evolve to package both. Our observation that the mi-

nus strand lacks the patterns of connectivity found in the plus strand aligns with this expectation.

These findings highlight a fundamental but under reported property of RNA virus genomes: in addition to encoding the structures and biological functionality of the plus strands, their sequences also determine the structures of the minus strands. Importantly, these structures, and their corresponding functions, need not be similar. Understanding these differences may help reveal trade-offs that viruses navigate, such as balancing stability, packaging efficiency, and replication dynamics. Addressing these questions requires tools that enable direct structural comparisons between plus and minus strands. Our patch–probe method provides one such tool.

Summary of results on STMV RNA

Below we summarize our results on STMV RNA, making direct comparisons to prior structural studies by Athavale *et al.* [42], Archer *et al.* [44], Larman *et al.* [45], and Schroeder *et al.* [46, 48]:

1. The signal map in Fig. 2 provides a quantitative description of the RNA connectivity, which reflects the full ensemble of structures in solution. In contrast, prior studies focused primarily on the most-probable secondary structure.

2. The map of dominant signals in Fig. 3 agrees with the consensus T-shaped domain established by prior studies, and thus supports the globally folded structures reported by Athavale *et al.*, Archer *et al.*, and Larman *et al.*, and contradicts the purely locally folded structure reported by Schroeder *et al.*

3. The dominant signals in Fig. 3 also address regions of the molecule where there is no consensus. Remarkably, the map reflects dominant connections that are individually consistent with unique features in each of the prior structures, including the structure by Schroeder *et al.* This result suggests that the population of RNA molecules adopts a distribution of structures, and that prior studies reflect only a fraction of that distribution.

4. The prevalence maps in Fig. 4 reveal changes in long-range connectivity caused by replacing uridine with pseudouridine in the RNA sequence. These changes can be visualized directly from the two-dimensional prevalence data, without reference to any specific structural model, something not possible with prior one-dimensional probing data.

5. The signal and prevalence maps for the minus strand in Fig. 5 reveal qualitative differences in connectivity compared to the plus strand, including fewer long-range connections.

Addressing longer RNA molecules

Because the patch–probe method is two-dimensional and the number of coupling-signal measurements increases quadratically with the RNA length, mapping the connectivity of very long RNA sequences can require more microarray spots than are available on a single array. In principle, multiple arrays can be combined to map RNA molecules of arbitrary length, since the prevelence values f_{ij} can be directly compared across experiments. Another approach is to reduce the resolution of the map by increasing the patch length and the spacing between 5′-ends of adjacent probes. Reasonable modifications to the patch length and probe spacing enable the study of RNA molecules of lengths up to 10 000 nucleotides using a single array, as discussed in Supplementary data and Supplementary Fig. S3. To facilitate the design process, we have included a Jupyter Notebook that automatically generates microarray layouts and patch sequences, given an input RNA sequence and a desired patch length and probe spacing (see Supplementary data).

Validating specific connections in arbitrarily long RNA molecules

Not all studies require comprehensive 2D mapping. For example, when partial information about a specific long-range connection already exists, be it from folding models or previous experiments, one may be interested simply in verifying whether or not that specific connection appears to a significant extent in the population. In such cases, a scaled-down version of the patch–probe method can be applied: one can check whether a putative connection is prevalent by binding a single patch to one endpoint of a putative connection and then measuring the coupling signals at the other potential endpoints. Because this approach does not involve a full set of patches, it can be applied to arbitrarily long RNA molecules, using only a small fraction of the array.

We demonstrate this approach by applying it to a well-studied long-range connection in the 3569-nt RNA genome of bacteriophage MS2. Thermodynamic folding models [60], mutational analysis [61], and X-ray footprinting measurements [62] have all suggested a long-range connection, termed the Min Jou interaction [63], between regions 1427–1433 and 1738–1744 of MS2 RNA. However, recent cryo-electron microscopy studies of infectious MS2 particles have not directly resolved this interaction within the packaged RNA genome [64, 65]. To test whether the Min Jou interaction is present in MS2 RNA transcripts, we apply a single 20-nt patch to nucleotides 1427–1446, which corresponds to the 5′-endpoint of the putative connection (Fig. 6A), and then measure coupling signals with a complete set of 12-nt probes (Fig. 6B). We observe large signals downstream of the patch site, contained within the region defined by the Min Jou interaction (Fig. 6B), providing strong support for this putative connection in our *in vitro* transcripts.

In addition to the peak centered at nucleotide 1745 that is consistent with the Min Jou interaction, we observe other peaks consistent with coexisting connections (Fig. 6). The largest of these peaks arise directly downstream of the patch site (Fig. 6B) and likely correspond to short-range connections that have been disrupted by the portion of the patch that extends beyond the Min Jou interaction (which we added to ensure strong binding between the patch and the RNA) (Fig. 6A). In contrast to the multiple peaks observed with STMV RNA (Figs 3A and 3D, ellipses 4 and 5), which reflect connections that do not coexist within the same structure, the multiple peaks observed with MS2 likely represent connections that are nested between the endpoints of the Min Jou interaction: thermodynamic folding calculations performed using RNAsubopt [16] reveal that all of the suboptimal structures containing the Min Jou interaction also contain nested connections consistent with the structure shown in Fig. 6A. How a single patch generates multiple peaks over several hundred nucleotides of the RNA sequence is an open question. One possibility is that the Min Jou interaction stabilizes the connections nested between its endpoints, such that disrupting the interaction weakens the nested connections and increases their binding to probes.

Discussion

Throughout, we have been careful to distinguish between connectivity and structure. Two-dimensional methods—including multidimensional chemical probing, proximity ligation, and the patch–probe method described here—measure intramolecular connections. If there were only one secondary structure, then the connections measured by these methods would determine that structure to within the resolution of the technique. But if there are variations in secondary structure within the population of molecules, these methods reveal only the connections that are present in the population. They do not reveal how those connections are grouped together into structures nor how many different structures exist, though such information could be gleaned through additional modeling and assumptions [66–68].

Nonetheless, information about the connectivity of long RNA molecules can provide important clues about their biological function. For example, the RNA genomes of certain viruses form connections that are thought to direct a range of functions, including the production of viral proteins [69] and the replication of new viral RNA strands [70], and there is growing evidence that these connections can rearrange in response to changing conditions [71–74], possibly triggering changes in functionality. Quantifying the prevalence of connections is therefore an important step in understanding how RNA virus genomes orchestrate infections, and could inform strategies for blocking infections by pathogenic viruses.

We use the patch–probe method to infer the prevalence of connections within STMV RNA, a well-established model system for RNA structure studies. Plus strands of the STMV genome have been extensively characterized using chemical probing, computational modeling, and direct imaging, providing a robust benchmark for validating our approach. Indeed, we observe long-range connections consistent with the consensus T-shaped domain found in previous studies. At the
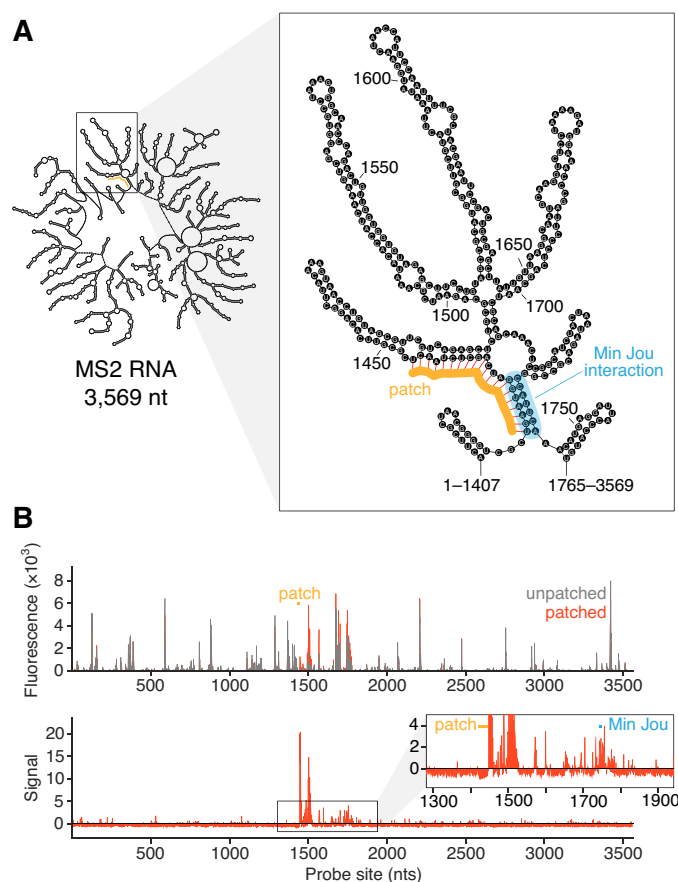
Figure 6. Validation of a long-range connection in a 3569-nt RNA molecule. (**A**) A secondary structure model of bacteriophage MS2 RNA derived from a combination of electron microscopy data and RNA folding calculations [65]. We bind a DNA patch to MS2 RNA at nucleotides 1427–1446, the 5′-endpoint of the Min Jou interaction [63]. (**B**) We observe changes in probe binding downstream of the patch site within the region bounded by the 3′-endpoint of the Min Jou interaction, corroborating the presence of this long-range connection in our *in vitro* transcripts.

same time, STMV RNA offers opportunities for new discovery: regions outside of the T-domain remain poorly resolved, and previous studies have produced conflicting structural interpretations. Our measurements help clarify these discrepancies by revealing that multiple distinct structures can coexist within the population. Furthermore, we used the method to detect structural changes induced by nucleotide modifications. Replacing uridine with pseudouridine alters the prevalence of specific long-range connections, demonstrating that chemical modifications can shift the RNA structural ensemble. Finally, by analyzing the previously uncharacterized minus strands, we uncover qualitative differences in connectivity that are not accessible through standard one-dimensional chemical probing. These results were obtained without reference to folding models of any kind, highlighting a key advantage of the patch–probe method.

Multidimensional chemical probing and proximity ligation techniques, such as mutate-and-map [18], Pair-MaP [23], PARIS [27], and COMRADES [30], can also detect competing connections without folding models, and we therefore consider how the patch–probe method compares to these techniques. Chemical probing and proximity ligation have the major advantage that they can address multiple sequences in parallel and *in vivo*, whereas the patch–probe method currently only works with a single sequence *in vitro* or, potentially, in cell extract. Furthermore, the signals measured in a proximity ligation experiment arise directly from RNA connections,

whereas the signals in chemical probing and patch–probe experiments reflect perturbations to the RNA connectivity and are therefore less direct. However, inferring the prevalence of connections from proximity ligation data is not always straightforward. In principle, the magnitude of the ligation signal should scale with prevalence, but in practice some steps of the protocol, including ligation and PCR, can bias the measured signals in ways that are difficult to correct for [26, 67]. Inferring the prevalence from patch–probe data is simpler: after correcting for the patch and probe affinities, the magnitude of each coupling peak provides a quantitative measure of the prevalence.

To interpret the prevalence values we assume that each coupling peak corresponds to a pairwise connection in the RNA molecules. While the approximate symmetry of the 2D data (Figs. 1F, 2, 3A, and 4) suggests that this assumption holds for many of the couplings measured in the experiment, there are scenarios in which the assumption might not hold, which could lead to the identification of spurious connections. For example, in regions of the RNA sequence directly flanking the patch site (Fig. 6), binding of the patch oligonucleotide could affect nearby probe binding even if those regions are not directly connected to the patch site. Furthermore, patch binding could induce nonlocal rearrangements of the RNA connectivity, involving cascades of multiple connections.

These scenarios highlight an important feature of the method: DNA oligonucleotides do not merely report on RNA

structure; they can also modify it. The ability to modify and interact with the folded structure of an RNA molecule could enable new ways of measuring collective aspects of its folding process. Much as how DNA origami uses many staple oligonucleotides to direct a long DNA molecule to fold into prescribed structures [75], it is possible that multiple patches could be used to drive a long RNA molecule away from its native folds in order to address larger regions of the folding landscape. For example, we imagine perturbing the RNA molecule in a way that mimics the unfolding and refolding effects of cellular enzymes during biological processes like transcription.

Finally, as an alternative to mapping the connectivity of an RNA molecule across its entire sequence, we show that the patch–probe method can be scaled down to study specific pairwise connections. This ability to focus on specific connections sets the method apart from other 2D techniques, offering a practical and efficient means of validating predicted long-range interactions. We believe this focused approach can complement existing structure determination methods, enhancing our collective toolkit for exploring RNA structure.

Acknowledgements

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

A US patent application (application serial number 17/482765) has been filed on the patch–probe binding method for determining the secondary structure of a nucleic acid by R.F.G., T.K.C., V.N.M., O.K., and M.P.B.

Funding

Data availability

Raw microarray scan images, integrated fluorescence values, and MCMC samples from Bayesian inference have been deposited in the Harvard Dataverse under a CC-0 license at https://doi.org/10.7910/DVN/83HMGQ. The code used to extract fluorescence values, infer coupling signals and prevalences, and design patch oligonucleotides and microarray layouts is available at https://github.com/manoharan-lab/patch-probe-code under the GNU General Public License v3 or later. The code used in this publication is archived at Zenodo as "manoharan-lab/patch-probe-code: initial release (v1.0)", https://doi.org/10.5281/zenodo.15364868.

References

1. Yoffe AM, Prinsen P, Gopal A *et al.* Predicting the sizes of large RNA molecules. *Proc Natl Acad Sci USA* 2008;**105**:16153–8. https://doi.org/10.1073/pnas.0808089105

2. Gopal A, Egecioglu DE, Yoffe AM *et al.* Viral RNAs are unusually compact. *PLoS One* 2014;**9**:e105875. https://doi.org/10.1371/journal.pone.0105875

3. Tubiana L, Božič AL, Micheletti C *et al.* Synonymous mutations reduce genome compactness in icosahedral ssRNA viruses. *Biophys J* 2015;**108**:194–202. https://doi.org/10.1016/j.bpj.2014.10.070

4. Berkhout B, van Duin J. Mechanism of translational coupling between coat protein and replicase genes of RNA bacteriophage MS2. *Nucleic Acids Res* 1985;**13**:6955–67. https://doi.org/10.1093/nar/13.19.6955

5. Huthoff H, Berkhout B. Two alternating structures of the HIV-1 leader RNA. *RNA* 2001;**7**:143–57. https://doi.org/10.1017/S1355838201001881

6. Ding Y, Lawrence CE. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res* 2003;**31**:7280–301. https://doi.org/10.1093/nar/gkg938

7. Wu B, Grigull J, Ore MO *et al.* Global organization of a positive-strand RNA virus genome. *PLoS Pathog* 2013;**9**:e1003363. https://doi.org/10.1371/journal.ppat.1003363

8. Nicholson BL, White KA. Functional long-range RNA–RNA interactions in positive-strand RNA viruses. *Nat Rev Microbiol* 2014;**12**:493–504. https://doi.org/10.1038/nrmicro3288

9. Cate JH, Gooding AR, Podell E *et al.* Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science* 1996;**273**:1678–85. https://doi.org/10.1126/science.273.5282.1678

10. Bonilla SL, Kieft JS. The promise of cryo-EM to explore RNA structural dynamics. *J Mol Biol* 2022;**434**:167802. https://doi.org/10.1016/j.jmb.2022.167802

11. Merino EJ, Wilkinson KA, Coughlan JL *et al.* RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). *J Am Chem Soc* 2005;**127**:4223–31. https://doi.org/10.1021/ja043822v

12. Tijerina P, Mohr S, Russell R. DMS footprinting of structured RNAs and RNA–protein complexes. *Nat Protoc* 2007;**2**:2608–23. https://doi.org/10.1038/nprot.2007.380

13. Eddy SR. Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. *Annu Rev Biophys* 2014;**43**:433–56. https://doi.org/10.1146/annurev-biophys-051013-022950

14. Deigan KE, Li TW, Mathews DH *et al.* Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci USA* 2009;**106**:97–102. https://doi.org/10.1073/pnas.0806929106

15. Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* 2010;**11**:129. https://doi.org/10.1186/1471-2105-11-129

16. Lorenz R, Bernhart SH, Höner zu Siederdissen C *et al.* ViennaRNA package 2.0. *Algorithm Mol Biol* 2011;**6**: 26. https://doi.org/10.1186/1748-7188-6-26

17. Amman F, Bernhart SH, Doose G *et al.* The trouble with long-range base pairs in RNA folding. In: Setubal JC, Almeida NF (eds.), *Advances in Bioinformatics and Computational Biology. Lecture Notes in Computer Science*. Cham: Springer International Publishing, 2013, 1–11. https://doi.org/10.1007/978-3-319-02624-4_1

18. Kladwang W, VanLang CC, Cordero P *et al.* A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nat Chem* 2011;**3**:954–62. https://doi.org/10.1038/nchem.1176

19. Petzold K, Al-Hashimi HM. Adding a second dimension. *Nat Chem* 2011;**3**:913–5. https://doi.org/10.1038/nchem.1209

20. Cheng CY, Fang-Chieh C, Wipapat K *et al.* Consistent global structures of complex RNA states through multidimensional chemical mapping. *eLife* 2015;**4**: e07600. https://doi.org/10.7554/eLife.07600

21. Cheng CY, Kladwang W, Yesselman JD *et al.* RNA structure inference through chemical mapping after accidental or intentional mutations. *Proc Natl Acad Sci USA* 2017;**114**:9876–81. https://doi.org/10.1073/pnas.1619897114

22. Homan PJ, Favorov OV, Lavender CA *et al.* Single-molecule correlated chemical probing of RNA. *Proc Natl Acad Sci USA* 2014;**111**:13858–63. https://doi.org/10.1073/pnas.1407306111

23. Mustoe AM, Lama NN, Irving PS *et al.* RNA base-pairing complexity in living cells visualized by correlated chemical probing. *Proc Natl Acad Sci USA* 2019;**116**:24574–82. https://doi.org/10.1073/pnas.1905491116

24. Olson SW, Turner AMW, Arney JW *et al.* Discovery of a large-scale, cell-state-responsive allosteric switch in the 7SK RNA using DANCE-MaP. *Mol Cell* 2022;**82**:1708–23. https://doi.org/10.1016/j.molcel.2022.02.009

25. Mustoe AM, Weidmann CA, Weeks KM. Single-molecule correlated chemical probing: a revolution in RNA structure analysis. *Accounts Chem Res* 2023;**56**:763–75. https://doi.org/10.1021/acs.accounts.2c00782

26. Ramani V, Qiu R, Shendure J. High-throughput determination of RNA structure by proximity ligation. *Nat Biotechnol* 2015;**33**:980–84. https://doi.org/10.1038/nbt.3289

27. Lu Z, Zhang QC, Lee B *et al.* RNA duplex map in living cells reveals higher-order transcriptome structure. *Cell* 2016;**165**:1267–79. https://doi.org/10.1016/j.cell.2016.04.028

28. Aw JGA, Shen Y, Wilm A *et al. In vivo* mapping of eukaryotic RNA interactomes reveals principles of higher-order organization and regulation. *Mol Cell* 2016;**62**:603–17. https://doi.org/10.1016/j.molcel.2016.04.028

29. Weidmann CA, Mustoe AM, Weeks KM. Direct duplex detection: an emerging tool in the RNA structure analysis toolbox. *Trends Biochem Sci* 2016;**41**:734–36. https://doi.org/10.1016/j.tibs.2016.07.001

30. Ziv O, Gabryelska MM, Lun ATL *et al.* COMRADES determines *in vivo* RNA structures and interactions. *Nat Methods* 2018;**15**:785–8. https://doi.org/10.1038/s41592-018-0121-0

31. Ziv O, Price J, Shalamova L *et al.* The short- and long-range RNA–RNA interactome of SARS-CoV-2. *Mol Cell* 2020;**80**:1067–77. https://doi.org/10.1016/j.molcel.2020.11.004

32. Uhlenbeck OC, Baller J, Doty P. Complementary oligonucleotide binding to the anticodon loop of Fmet-transfer RNA. *Nature* 1970;**225**:508–10. https://doi.org/10.1038/225508a0

33. Lewis JB, Doty P. Derivation of the secondary structure of 5S RNA from its binding of complementary oligonucleotides. *Nature* 1970;**225**:510–2. https://doi.org/10.1038/225510a0

34. Sohail M, Akhtar S, Southern EM. The folding of large RNAs studied by hybridization to arrays of complementary oligonucleotides. *RNA* 1999;**5**:646–55. https://doi.org/10.1017/S1355838299982195

35. Walton SP, Stephanopoulos GN, Yarmush ML *et al.* Thermodynamic and kinetic characterization of antisense oligodeoxynucleotide binding to a structured mRNA. *Biophys J* 2002;**82**:366–77. https://doi.org/10.1016/S0006-3495(02)75401-5

36. Kierzek E, Kierzek R, Turner DH *et al.* Facilitating RNA structure prediction with microarrays. *Biochemistry* 2006;**45**:581–93. https://doi.org/10.1021/bi051409+

37. Liang R, Kierzek E, Kierzek R *et al.* Comparisons between chemical mapping and binding to isoenergetic oligonucleotide microarrays reveal unexpected patterns of binding to the bacillus subtilis RNase P RNA specificity domain. *Biochemistry* 2010;**49**:8155–68. https://doi.org/10.1021/bi100286n

38. Lukasiewicz AJ, Contreras LM. Antisense probing of dynamic RNA structures. *Methods* 2020;**183**:76–83. https://doi.org/10.1016/j.ymeth.2020.01.015

39. Kaplinski L, Scheler O, Parkel S *et al.* Detection of tmRNA molecules on microarrays at low temperatures using helper oligonucleotides. *BMC Biotechnology* 2010;**10**:34. https://doi.org/10.1186/1472-6750-10-34

40. Buvoli A, Buvoli M, Leinwand LA. Enhanced detection of tRNA isoacceptors by combinatorial oligonucleotide hybridization. *RNA* 2000;**6**:912–8. https://doi.org/10.1017/S1355838200000339

41. Southern EM. DNA microarrays. In: Rampal JB (ed.), *DNA Arrays: Methods and Protocols. Methods in Molecular Biology*. Totowa, NJ: Humana Press, 2001, 1–15. https://doi.org/10.1385/1-59259-234-1:1

42. Athavale SS, Gossett JJ, Bowman JC *et al. In vitro* secondary structure of the genomic RNA of satellite tobacco mosaic virus. *PLoS One* 2013;**8**:e54384. https://doi.org/10.1371/journal.pone.0054384

43. Shaklee PN. Negative-strand RNA replication by Qβ and MS2 positive-strand RNA phage replicases. *Virology* 1990;**178**:340–3. https://doi.org/10.1016/0042-6822(90)90417-P

44. Archer EJ, Simpson MA, Watts NJ *et al.* Long-range architecture in a viral RNA genome. *Biochemistry* 2013;**52**:3182–90. https://doi.org/10.1021/bi4001535

45. Larman BC, Dethoff EA, Weeks KM. Packaged and free satellite tobacco mosaic virus (STMV) RNA genomes adopt distinct conformational states. *Biochemistry* 2017;**56**:2175–83. https://doi.org/10.1021/acs.biochem.6b01166

46. Schroeder SJ, Stone JW, Bleckley S *et al.* Ensemble of secondary structures for encapsidated satellite tobacco mosaic virus RNA consistent with chemical probing and crystallography constraints. *Biophys J* 2011;**101**:167–75. https://doi.org/10.1016/j.bpj.2011.05.053

47. Zeng Y, Larson SB, Heitsch CE *et al.* A model for the structure of satellite tobacco mosaic virus. *J Struct Biol* 2012;**180**:110–6. https://doi.org/10.1016/j.jsb.2012.06.008

48. Schroeder SJ. Probing viral genomic structure: alternative viewpoints and alternative structures for satellite tobacco mosaic virus RNA. *Biochemistry* 2014;**53**:6728–37. https://doi.org/10.1021/bi501051k

49. Garmann RF, Gopal A, Athavale SS *et al.* Visualizing the global secondary structure of a viral RNA genome with cryo-electron

microscopy. *RNA* 2015;**21**:877–86. https://doi.org/10.1261/rna.047506.114

50. Charette M, Gray MW. Pseudouridine in RNA: what, where, how, and why. *IUBMB Life* 2000;**49**:341–51. https://doi.org/10.1080/152165400410182

51. Karikó K, Muramatsu H, Welsh FA *et al.* Incorporation of pseudouridine into mRNA yields superior nonimmunogenic vector with increased translational capacity and biological stability. *Mol Ther* 2008;**16**:1833–40. https://doi.org/10.1038/mt.2008.200

52. Hudson GA, Bloomingdale RJ, Znosko BM. Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligoribonucleotides. *RNA* 2013;**19**:1474–82. https://doi.org/10.1261/rna.039610.113

53. Kierzek E, Malgowska M, Lisowiec J *et al.* The contribution of pseudouridine to stabilities and structure of RNAs. *Nucleic Acids Res* 2014;**42**:3492–501. https://doi.org/10.1093/nar/gkt1330

54. Dreher TW. Role of tRNA-like structures in controlling plant virus replication. *Virus Res* 2009;**139**:217–29. https://doi.org/10.1016/j.virusres.2008.06.010

55. Belkum Av, Abrahams JP, Pleij CWA *et al.* Five pseudoknots are present at the 204 nucleotides long 3′ noncoding region of tobacco mosak virus RNA. *Nucleic Acids Res* 1985;**13**:7673–86. https://doi.org/10.1093/nar/13.21.7673

56. Routh G, Yassi MNA, Rao A *et al.* Replication of wild-type and mutant clones of satellite tobacco mosaic virus in *Nicotiana benthamiana* protoplasts. *J Gen Virol* 1997;**78**:1271–5. https://doi.org/10.1099/0022-1317-78-6-1271

57. Diener T. Viroids: structure and function. *Science* 1979;**205**:859–66. https://doi.org/10.1126/science.472709

58. Larson SB, Koszelak S, Day J *et al.* Three-dimensional structure of satellite tobacco mosaic virus at 2.9 Å resolution. *J Mol Biol* 1993;**231**:375–91. https://doi.org/10.1006/jmbi.1993.1289

59. Larson SB, McPherson A. Satellite tobacco mosaic virus RNA: structure and implications for assembly. *Curr Opin Struct Biol* 2001;**11**:59–65. https://doi.org/10.1016/S0959-440X(00)00166-4

60. Fiers W, Contreras R, Duerinck F *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* 1976;**260**:500–7. https://doi.org/10.1038/260500a0

61. Licis N, van Duin J. Structural constraints and mutational bias in the evolutionary restoration of a severe deletion in RNA phage MS2. *J Mol Evol* 2006;**63**:314–29. https://doi.org/10.1007/s00239-005-0012-8

62. Chandler-Bostock R, Bingham RJ, Clark S *et al.* Genome-regulated assembly of a ssRNA virus may also prepare it for infection. *J Mol Biol* 2022;**434**:167797. https://doi.org/10.1016/j.jmb.2022.167797

63. Jou WM, Haegeman G, Ysebaert M *et al.* Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* 1972;**237**:82–8. https://doi.org/10.1038/237082a0

64. Koning RI, Gomez-Blanco J, Akopjana I *et al.* Asymmetric cryo-EM reconstruction of phage MS2 reveals genome structure *in situ*. *Nat Commun* 2016;**7**:12524. https://doi.org/10.1038/ncomms12524

65. Dai X, Li Z, Lai M *et al. In situ* structures of the genome and genome-delivery apparatus in a single-stranded RNA virus. *Nature* 2017;**541**:112–6. https://doi.org/10.1038/nature20589

66. Zhou J, Li P, Zeng W *et al.* IRIS: a method for predicting *in vivo* RNA secondary structures using PARIS data. *Quant Biol* 2020;**8**:369–81. https://doi.org/10.1007/s40484-020-0223-4

67. Kudla G, Wan Y, Helwak A. RNA conformation capture by proximity ligation. *Annu Rev Genom Hum Genet* 2020;**21**:81–100. https://doi.org/10.1146/annurev-genom-120219-073756

68. Van Damme R, Li K, Zhang M *et al.* Chemical reversible crosslinking enables measurement of RNA 3D distances and alternative conformations in cells. *Nat Commun* 2022;**13**:911. https://doi.org/10.1038/s41467-022-28602-3

69. Groeneveld H, Thimon K, Duin Jv. Translational control of maturation-protein synthesis in phage MS2: a role for the kinetics of RNA folding? *RNA* 1995;**1**:79–88.

70. Filomatori CV, Lodeiro MF, Alvarez DE *et al.* A 5′ RNA element promotes dengue virus RNA synthesis on a circular genome. *Genes Dev* 2006;**20**:2238–49. https://doi.org/10.1101/gad.1444206

71. Huber RG, Lim XN, Ng WC *et al.* Structure mapping of dengue and Zika viruses reveals functional long-range interactions. *Nat Commun* 2019;**10**:1408. https://doi.org/10.1038/s41467-019-09391-8

72. Sage VL, Kanarek JP, Snyder DJ *et al.* Mapping of influenza virus RNA–RNA interactions reveals a flexible network. *Cell Rep* 2020;**31**:107823. https://doi.org/10.1016/j.celrep.2020.107823

73. Zhang Y, Huang K, Xie D *et al. In vivo* structure and dynamics of the SARS-CoV-2 RNA genome. *Nat Commun* 2021;**12**:5695. https://doi.org/10.1038/s41467-021-25999-1

74. Fairman CW, Lever AML, Kenyon JC. Evaluating RNA structural flexibility: viruses lead the way. *Viruses* 2021;**13**:2130. https://doi.org/10.3390/v13112130

75. Rothemund PWK. Folding DNA to create nanoscale shapes and patterns. *Nature* 2006;**440**:297–302. https://doi.org/10.1038/nature04586